대 한 민 국 특 허 청
KOREAN INDUSTRIAL PROPERTY OFFICE

별첨 사본은 아래 출원의 원본과 동일함을 증명함.

This is to certify that the following application annexed hereto is a true copy from the records of the Korean Industrial Property Office.

출 원 번 호 : 특허출원 2000년 제 50418 호
Application Number

출 원 년 월 일 : 2000년 08월 29일
Date of Application

출 원 인 : 한국과학기술원
Applicant(s)

CERTIFIED COPY OF PRIORITY DOCUMENT

2000 년 12 월 27 일

특 허 청
COMMISSIONER

【서류명】 특허출원서

【권리구분】 특허

【수신처】 특허청장

【제출일자】 2000.08.29

【발명의 명칭】 대용량 데이터베이스에서의 고속에 의한 다해상도의 최적 탐색방법

【발명의 영문명칭】 METHOD FOR OPTIMAL RETRIEVAL OF MULTI-RESOLUTION BY HIGH SPEED ON THE GREAT CAPACITY DATABASE

【출원인】

【명칭】 한국과학기술원

【출원인코드】 3-1998-098866-1

【대리인】

【성명】 이종일

【대리인코드】 9-1998-000471-4

【포괄위임등록번호】 2000-039220-2

【대리인】

【성명】 조희연

【대리인코드】 9-2000-000220-0

【포괄위임등록번호】 2000-039231-8

【발명자】

【성명의 국문표기】 나종범

【성명의 영문표기】 NA,Jong Bum

【주민등록번호】 530523-1037313

【우편번호】 305-390

【주소】 대전광역시 유성구 전민동 464-1 엑스포아파트 404동 506호

【국적】 KR

【발명자】

【성명의 국문표기】 송병철

【성명의 영문표기】 SONG,Byung Chul

【주민등록번호】 721108-1446725

【우편번호】 305-701

【주소】 대전광역시 유성구 구성동 373-1 한국과학기술원 전자전산학과

【국적】 KR

【심사청구】 청구

【취지】 특허법 제42조의 규정에 의한 출원, 특허법 제60조의 규정에 의한 출원심사 를 청구합니다. 대리인
이종일 (인) 대리인
조희연 (인)

【수수료】

【기본출원료】	20	면	29,000 원
【가산출원료】	8	면	8,000 원
【우선권주장료】	0	건	0 원
【심사청구료】	11	항	461,000 원
【합계】	498,000 원		
【감면사유】	정부출연연구기관		
【감면후 수수료】	249,000 원		

【요약서】

【요약】

본 발명은 대용량 데이터베이스에서의 고속에 의한 다해상도의 최적 탐색방법에 관한 것이다. 종래의 클러스터 기반 기법들은 최적 검색을 보장하지 못하며, 탐색 속도가 만족 할만한 검색 정확도를 얻기에는 충분히 빠르지 못하였다.

이에 본 발명은 1) 가능성이 있는 클러스터와 가능성이 없는 클러스터를 정확하게 구분하는 부등식을 유도하고 이를 이용한 최적의 탐색 기법을 구현토록 한다. 2) 고속 처리를 위한 다해상도 데이터 구조에 기반한 부등식을 유도하고 이를 이용한 고속 최적 탐색 기법을 구현토록 한다.

본 발명에 따른 대용량 데이터베이스에서의 고속에 의한 다해상도의 최적 탐색방법은 1) 데이터베이스 내의 모든 데이터들을 일정 수의 클러스터(유사한 특징을 갖는 클러스터)들로 나눈다. 2) 각 클러스터와 문의자간 거리의 하계(lower bound)를 구하여 가능성이 없다고 판단될 경우 그 클러스터를 제거하고 최종적으로 가능성이 있다고 판단된 클러스터들의 데이터들 중에서 최적 정합자를 찾는다. 3) 보다 많은 계산량 감소를 위해 탐색 과정에서 불필요한 특징 정합 연산을 줄이기 위한 다해상도 데이터 구조에 기반한 거리부등식 성질을 유도한다.

【대표도】

도 2

【색인어】

데이터베이스, 다해상도, 클러스트, 전역탐색, 고속탐색

【명세서】

【발명의 명칭】

대용량 데이터베이스에서의 고속에 의한 다해상도의 최적 탐색방법{METHOD FOR OPTIMAL RETRIEVAL OF MULTI-RESOLUTION BY HIGH SPEED ON THE GREAT CAPACITY DATABASE}

【도면의 간단한 설명】

도 1은 종래 클러스터에 기반한 탐색 방법들의 본질적인 문제점을 설명하기 위한 도면

도 2는 본 발명에 따라 한 클러스터에 대한 거리의 부등식 특징을 설명하기 위한 도면

도면 3은 본 발명에 따라 2^L빈을 갖는 밝기 히스토그램 X의 다 해상도 데이터 구조를 나타낸 도면

도면 4는 본 발명에 따라 상위 M개의 최적 정합자들의 최소 거리 배열을 나타낸 모식도

도 5는 본 발명에 따라 최적 정합자들을 정확하게 못하게 되는 그릇된 판정의 예를 설명하기 위한 도면

【발명의 상세한 설명】

【발명의 목적】

【발명이 속하는 기술분야 및 그 분야의 종래기술】

<6> 본 발명은 대용량 데이터베이스에서의 고속에 의한 다해상도의 최적 탐색방법에 관한 것이다. 특히, 데이터베이스에서 매우 빠른 속도로 원하는 정보를 정확하고 빠르게

탐색할 수 있는 부등식을 유도하고, 이를 이용하여 고속으로 최적의 탐색방법을 구현토록 하는 기술에 관한 것이다.

<7> 유사도 측정자(similarity measure)에 따른 문의자(query)의 최적 정합자 (the best match)를 찾기 위해서는 데이터베이스 안의 모든 데이터들에 대해 전역 탐색을 수행하여야 한다. 통상의 전역 탐색기법(Exhaustive Search Algorithm : ESA)은 계산량이 많기 때문에 최근 여러 고속 탐색 기법들이 개발된 바 있다.

<8> Berman과 Shapiro는 탐색 과정에서 최적일 가능성이 전혀 없는 후보들을 적은 계산량으로 제거하기 위해 삼각형 부등식을 도입하였다. 이는 추가적인 계산량 감소를 위해 여러 가지 거리 측정자들과 대표 데이터들을 동시에 사용하기도 했지만, 탐색 속도가 대표 데이터들에 따라 변화가 크고, 대용량 데이터베이스에서 만족할만한 성능을 보이지 못했다.

<9> 최근 트라이앵글 트리(Triangle Trie)라는 데이터 구조를 적용하여 성능을 개선 하였지만 탐색 속도는 여전히 대표 데이터들의 트리 깊이, 문턱값 등에 많은 영향을 받는다.

<10> 한편, Krishnamachari과 Mottaleb는 데이터베이스에 있는 데이터들을 계층적 클러스터링 기법으로 비슷한 특징을 갖는 클러스터들로 분할하는 클러스터 기반 색인 기법을 새롭게 제안하였다.

<11> 이는 탐색 과정에서 문의자 데이터를 데이터베이스의 모든 데이터들과 비교하지 않고, 클러스터링에 의해 일부 데이터들과 비교하기 때문에 계산량을 현저히 줄일 수 있다.

<12> 또한, 클러스터 기반 기법들은 원하는 검색 정확도(retrieval accuracy)를 얻기 위한 비교 횟수가 데이터베이스 크기와 선형적으로 비례하지 않기 때문에 대용량 데이터베이스에 적합하다고 할 수 있다.

<13> 도 1은 종래 클러스터 기반에서의 탐색 기법들에 의해 발생되는 문제점을 나타낸 도면이다.

<14> 도 1에 도시된 바와 같이, 중심점들 중 C_2 가 문의자 Q에 가장 가깝기 때문에 클러스터(Cluster) 2는 후보로 선택된다. 클러스터 2에 속한 각 원소와 Q와의 거리를 계산함으로써 X_2를 최적 정합자로 선택한다. 그러나, 실제 최적 정합자는 클러스터 1의 X_8이다.

<15> 이런 문제가 발생하는 이유는 실제 최적 정합자가 속한 클러스터의 중심점이 항상 Q와 가장 가깝지 않기 때문이다. 따라서, Q와 가까운 여러 개의 클러스터들을 동시에 탐색하는 방법이 시도되었지만, 여전히 최적 검색은 보장하지 못했다.

<16> 또한, 기존 클러스터 기반 기법들은 최적 검색(optimal retrieval)을 보장하지 못하며, 탐색 속도가 만족 할만한 검색 정확도를 얻기에는 충분히 빠르지 않다는 단점이 있다.

【발명이 이루고자 하는 기술적 과제】

<17> 따라서, 본 발명은 상기한 문제점을 해결하기 위한 것으로서, 본 발명의 목적은 가능성이 있는 클러스터와 가능성이 없는 클러스터를 정확하게 구분하는 부등식을 유도하고 이를 이용한 최적의 탐색 기법을 구현토록 하는 대용량 데이터베이스에서의 고속에 의한 다해상도의 최적 탐색방법을 제공하는데 있다.

<18> 본 발명의 다른 목적은 고속 처리를 위한 다해상도 데이터 구조에 기반한 부등식을 유도하고 이를 이용한 고속 최적 탐색 기법을 구현토록 하는 대용량 데이터베이스에서의 고속에 의한 다해상도의 최적 탐색방법을 제공하는데 있다.

<19> 상기한 본 발명의 목적을 달성하기 위한 기술적 사상으로써 본 발명에 따른 대용량 데이터베이스에서의 고속에 의한 다해상도의 최적 탐색방법은 1) 데이터베이스 내의 모든 데이터들을 일정 수의 클러스터(유사한 특징을 갖는 클러스터)들로 나눈다. 2) 각 클러스터와 문의자간 거리의 하계(lower bound)를 구하여 가능성이 없다고 판단될 경우 그 클러스터를 제거하고 최종적으로 가능성이 있다고 판단된 클러스터들의 데이터들 중에서 최적 정합자를 찾는다. 3) 보다 많은 계산량 감소를 위해 탐색 과정에서 불필요한 특징 정합 연산을 줄이기 위한 다해상도 데이터 구조에 기반한 거리부등식 성질을 유도한다. 따라서, 단일 최적 정합자 뿐만 아니라 다수의 상위 최적 정합자들도 정확하게 찾을 수 있게 된다.

【발명의 구성 및 작용】

<20> 이하, 본 발명의 실시예에 대한 구성 및 그 작용을 첨부한 도면을 참조하면서 상세히 설명하기로 한다.

<21> 도 2는 본 발명에 따른 한 클러스터에 대한 거리 부등식 특징을 나타낸 모식도이다.

<22> 본 발명에 따른 한 클러스터에 대한 거리 부등식 특징을 설명하기 전에 데이터베이스의 클러스터링 과정을 살펴보면 다음과 같다.

<23> 먼저, 비슷한 특징을 갖는 데이터들이 하나의 클러스터를 이루도록 MacQueen K-평

균 클러스터링을 이용하여 정해진 수만큼의 클러스터들로 데이터베이스를 분할한다.

<24> 여기서, 사용할 수 있는 특징을 예로 들면, 영상 데이터의 경우는 색깔, 텍스쳐, 윤곽선 같은 정보이며, 음성 데이터의 경우는 pitch같은 정보가 가능하다. *K*개의 클러스터들은 각자의 평균 중심점을 갖는다.

<25> 클러스터링을 위한 연산은 실제 탐색과는 무관하게 이루어지므로 클러스터링에 걸리는 시간은 탐색 시간에 포함되지 않는다. 데이터베이스의 클러스터링은 다음과 같이 이루어진다.

<26> 단계 1) 클러스터의 갯수 (*K*; *K*<*N*)를 정한다.

<27> 단계 2) 클러스터 중심점들의 특징들, C_1, C_2.... C_k 을 초기화 한다. 데이터베이스 내 *K*개 데이터들을 임의로 선택한다. 효율적으로 초기화하기 위해 두 중심점들 간 최소 거리가 문턱값 보다 작지 않게 한다.

<28> 단계 3) 초기 중심점들로 선택된 데이터들을 제외한 나머지 데이터들 각각에 대해 가장 가까운 클러스터 중심점을 찾아 해당 클러스터에 포함시키고, 다음 수학식에 의해 각 중심점을 갱신한다.

<29> 【수학식 1】

$$C_\kappa=\frac{1}{n(\Phi_\kappa)+1}[n(\Phi_\kappa)C_\kappa+X_i]$$

<30> 여기서, X_i는 클러스터 *K*에 추가될 *i*번째 원소이며, Φ_k는 클러스터*k* 를, $n(\Phi_k)$는 Φ_k에 속한 원소들의 갯수 이다.

<31> 단계 4) 모든 원소들에 대해 수렴할 때까지 상기 단계 3을 반복 한다. 최종적으로 중심점 집합 $\Pi^0=\{C_1, C_2, \cdots, C_k\}$ 이 얻어진다.

<32> 도 2를 참조하여, 종래의 도 1에 도시된 클러스트에 기반에서 발생되는 본질적인 문제점의 해결 방안을 제시한다.

<33> 먼저, Q와 가장 가까운 클러스터에서 초기 최소 거리 $d_{\min,0}$ 를 구한다.

<34> 【수학식 2】

$$d_{\min,o} = \min_{X_i \in k_{\min}} d(X_i, Q)$$

<35> 상기 수학식 2에서

<36> 【수학식 3】

$$C_{k_{\min}} = \arg \min_{C_k \in \prod^0} d(C_k, Q)$$

<37> d(X,Y)는 두 특징 X와 Y의 L_1-놈 거리를 의미한다. 종래의 도면 1에서 $d_{\min,0}$는 d(X2,Q)이다. 여기서, $\Phi_{k_{\min}}$를 제외한 각 클러스터 내에서 중심점과 가장 먼 원소를 찾고, 그 거리를 다음과 같이 정의한다.

<38> 【수학식 4】

$$\delta_k = \max_{X_i \in \Phi_k} d(X_i, C_k)$$

<39> 각 클러스터의 δ는 미리 계산되어 저장된다. $d_{\min,0}$와 δ_k값들을 이용하여 각 클러스터가 최적 검색을 위해 탐색될 필요가 있는지 없는지를 판단한다. 이 결정을 위한 성질 1은 다음과 같다.

<40> 【수학식 5】

$$If d(C_k, Q) - \delta_k > d_{\min}, \min_{X_i \in \Phi_k} d(X_i, Q) > d_{\min}$$

<41> 성질 1이 기재되어 있는 수학식 5의 증명은 아래와 같다.

<42> 【수학식 6】

$$X_{i_{\min}} = \arg \min_{X_i \in \Phi_k} d(X_i, Q)$$

<43> 삼각형 부등식에 의해,

<44> 【수학식 7】

$$d(X_{i_{\min}}, Q) \geq d(C_k, Q) - d(X_{i_{\min}}, C_k)$$

<45> 수학식 4로부터,

<46> 【수학식 8】

$$\delta_k = \max_{X_i \in \Phi_k} d(X_i, C_k) \geq d(X_{i_{\min}}, C_k)$$

<47> 다음의 부등식이 수학식 7과 8로부터 얻어진다.

<48> 【수학식 9】

$$d(X_{i_{\min}}, Q) \geq d(C_k, Q) - d(X_{i_{\min}}, C_k) \geq d(C_k, Q) - \delta_k$$

<49> 만약 $d(c_k, Q) - \delta_k > d_{\min}$이면,

<50> 【수학식 10】

$$d(X_{i_{\min}}, Q) = \min_{X_{i_{\min}} \in \Phi_k} d(X_i, Q) > d_{\min}$$

<51> 따라서, 수학식 5가 성립됨을 증명할 수 있다.

<52> 상기와 같이 성질 1에서, $d(C_k, Q) - \delta_k$은 Q와 클러스터 k내 원소 간 거리의 하계(lower bound)를 의미한다.

<53> 만약 $d(C_k, Q) - \delta_k$이 $d_{\min}$보다 크면, 클러스터 k에는 $d_{\min}$보다 작은 거리를 갖는 원소가 존재할 수 없으므로, 클러스터 k는 더 이상 고려할 필요 없다.

<54> 따라서, 성질 1을 적용하면 효과적으로 가능성이 전혀 없는 모든 클러스터들을 정

확하게 제거할 수 있다.

<55> 그러나, 가장 가까운 클러스터를 찾는 과정과 최적 정합자를 찾는 과정은 여전히 상당한 계산량을 요구한다. 이 계산량을 줄이기 위해 다 해상도 데이터 구조에 바탕을 둔 또다른 부등식 성질을 유도하고, 이를 이용한 고속에 의한 최적의 탐색방법을 제안한다.

<56> 도 3은 본 발명에 따라 2^L빈을 갖는 밝기 히스토그램 X의 다 해상도 데이터 구조를 나타낸 모식도이다.

<57> 도 3에 도시된 다 해상도 데이터 구조를 살펴보면, 편의상 B (B=2^L)빈을 갖는 정규화된 밝기 히스토그램이 특징이라고 가정한다. 히스토그램 X의 다 해상도 데이터 구조는 히스토그램열 $\{X^0,\cdots,X^l,\cdots,X^L\}$로 정의될 수 있다.

<58> 여기서 $X=X^l$이다. X^l는 2^l개의 빈들을 가지며, X^{l+1}로부터 1/2비율로 해상도를 줄임으로써 얻어진다.

<59> 현재 계층의 각 화소값은 인접한 상위 계층의 두 화소값을 더해 얻어진다. 즉, $X^l(m)$이 X^l의 m번째 빈 값이라 할 때, $X^l(m)$은 다음과 같이 얻어진다.

<60> 【수학식 11】

$$X^l(m)=X^{l+1}(2m-1)+X^{l+1}(2m),1\le m\le 2^l$$

<61> 이어서, 다 해상도 특징 공간에서의 고속 최적 검색을 위한 성질 2를 살펴보기로 한다.

<62> 【수학식 12】

$$d(X,Y)\cong d^L(X,Y)\ge d^{L-1}(X,Y)\ge\cdots\ge d^l(X,Y)\ge\cdots\ge d^l(X,Y)\ge d^0(X,Y)$$

<63> 여기서, $d^l(X,Y)$는 계층 l에서의 두 히스토그램 X와 Y의 L_1-놈 거리, 즉 $d(X^l, Y^l)$을 의미한다.

<64> 성질 2가 기재되어 있는 수학식 12의 증명은 아래와 같다.

<65> 계층 l+1에서 두 히스토그램 X와 Y의 L_1- 놈 차는 다음과 같이 구해진다.

<66> 【수학식 13】

$$d^{l+1}(X,Y)=\sum_{m=1}^{2^{l+1}} | X^{l+1}(m)- Y^{l+1}(m) |$$

<67>

$$=\sum_{m=1}^{2^l}(| X^{l+1}(2m-1)- Y^{l+1}(2m-1) | + | X^{l+1}(2m)- Y^{l+1}(2m) |)$$

<68> X^{l+1}와 Y^{l+1}는 모두 2^{l+1}개의 빈을 가지며, $X^{l+1}(m)$은 X^{l+1}의 m번째 빈의 값이다.

<69> 한편, 수학식 11을 사용하여 계층 l에서의 거리를 다음과 같이 표현할 수 있다.

<70> 【수학식 14】

$$d^l(X,Y)=\sum_{m=1}^{2^l} | X^l(m)- Y^l(m) |$$

<71>

$$=\sum_{m=1}^{2^l} | X^{l+1}(2m-1)- Y^{l+1}(2m-1)+X^{l+1}(2m)- Y^{l+1}(2m) |$$

<72> $|A|+|B| \geq |A+B|$ 이므로,

<73> 【수학식 15】

$$\sum_{m=1}^{2^l}(| X^{l+1}(2m-1)- Y^{l+1}(2m-1) | + | X^{l+1}(2m)- Y^{l+1}(2m) |)$$

<74>

$$\geq \sum_{m=1}^{2^l} | X^{l+1}(2m-1)- Y^{l+1}(2m-1)+X^{l+1}(2m)- Y^{l+1}(2m) |$$

<75> 수학식 13, 14, 15로부터,

<76> 【수학식 16】

$$d^{l+1}(X,Y)=\sum_{m=1}^{2^l}(\,|\,X^{l+1}(2m-1)-Y^{l+1}(2m-1)\,|+|\,X^{l+1}(2m)-Y^{l+1}(2m)\,|\,)$$

<77>

$$\geq\sum_{m=1}^{2^l}|\,X^{l+1}(2m-1)-Y^{l+1}(2m-1)+X^{l+1}(2m)-Y^{l+1}(2m)\,|$$

<78>

$$=d^{l}(X,Y)$$

<79> 상기의 수학식 16으로부터 수학식 12가 성립되므로써 성질 2를 증명할 수 있다.

<80> 상기의 성질 2는 $d^{l}(X,Y)$이 특정 값보다 크면, $d^{L}(X,Y)$은 항상 그 특정 값보다 큼을 의미한다.

<81> 상위 계층에서의 거리 계산이 하위 계층에서의 거리 계산보다 많은 연산을 필요로 하고, 하위 계층들에서 가능성이 전혀 없는 후보들을 많이 제거할 수 있기 때문에 이 성질을 탐색 과정에 적용하면 탐색을 위한 계산량을 상당히 줄일 수 있다.

<82> N은 데이터베이스 $I=\{I_1,\cdots,I_i,\cdots,I_N\}$의 데이터 갯수이며, $\Omega^0=\{X_1,\cdots,X_i,\cdots,X_N\}$는 데이터들의 특징 집합이라고 하자. 각 데이터의 다 해상도 특징은 미리 계산되어 저장되어 있다.

<83> 상기 성질 2에 기반한 고속 다 해상도 전역 탐색 기법 (MSA_S)이 다음과 같이 요약될 수 있다.

<84> 단계 1) 문의자 특징 Q의 다 해상도 구조를 구한다.

<85> 단계 2) 초기 $d_{\min}$을 무한대로 설정한다.

<86> 단계 3) i와 l는 모두 1로 한다.

<87> 단계 4) $l=L$이면, 단계 6으로 간다. 만약 i 가 N보다 크면, 단계 7로 간다.

<88> 단계 5) $d^l(X_i,Q)$을 구한다. $d^l(X_i,Q)$이 d_{min} 보다 크면, 현재 후보 X_i를 제거하고, i와 l 을 각각 i+1과 1로 갱신한다. 그렇지 않으면, l을 l+1로 갱신한 후 단계 3으로 간다.

<89> 단계 6) $d^L(X_i,Q)$이 d_{min} 보다 크면, 현재 후보 X_i를 제거한다. 그렇지 않으면, d_{min}을 $d^L(X_i,Q)$로 갱신한다. i와 l 을 각각 i+1과 1로 갱신한 후 단계 4로 간다.

<90> 단계 7) 최종 d_{min}을 갖는 데이터를 최적 정합자로 선택한다.

<91> 이상에서와 같이 데이터베이스에 있는 각 데이터의 다 해상도 특징은 미리 계산되어 저장된다. 그러나, 문의자 데이터의 다 해상도 특징은 탐색 시간동안 얻어져야 하므로 그 계산량을 고려 해야만 한다.

<92> 정규화된 밝기 히스토그램 특징을 예로 들면, 계층 수가 8이므로 다 해상도 히스토그램을 얻기 위해서는 단지 254번의 덧셈 연산만이 필요하다. 1회의 정합 과정을 위해 511번의 덧셈 연산과 256번의 절대값 연산이 소요됨을 감안할 때 다 해상도 밝기 히스토그램을 위한 계산량은 무시할 만하다.

<93> 한편, 다 해상도 히스토그램들을 저장하기 위한 추가 메모리가 필요하나, 히스토그램의 크기가 데이터 크기보다 훨씬 작기 때문에 추가적인 메모리 증가는 무시할 만하다. 다른 특징들에 대해서도 상황은 마찬가지이다.

<94> 이어서, 성질 1의 클러스터 최적 제거 조건과 성질 2에 기반한 MSA_S를 이용하여, 고속 최적 탐색을 위한 새로운 클러스터 기반에서의 다 해상도 탐색 기법 (Cluster-based Multi-resolution Search Algorithm; CMSA)에 대하여 설명하기로 한다.

<95> 먼저, 문의자가 주어지면 MSA_S를 통해 문의자와 가장 근접한 클러스터 중심점을 찾

고, 그 클러스터 내에서 초기 최적 정합자와의 거리 $d_{\min}$를 구한다.

<96> 그리고나서, 성질 1의 클러스터 제거 조건에 따라 가능성이 있다고 판단된 클러스터들에 MSA_S를 적용하여 최적 정합자(들)를 찾는다. 가장 가까운 클러스터 중심점을 찾을 때 MSA_S를 사용하기 때문에, 클러스터 제거 과정에서 모든 $d^L(C_k,Q)$값들이 존재하지는 않는다.

<97> 왜냐하면 $d^{l_k}(C_k,Q)$이 $d_{\min}$ 보다 크면, l_K 보다 큰 계층들에서의 거리들 {d}^{{l}_{k}}({C}_{k}, Q),..., {d}^{L}({C}_{k},Q) 이 계산되지 않기 때문이다.

<98> 결국, 수학식 5의 성질 1을 클러스터 제거 과정에 도입하기 위해 다시 $d^L(C_k,Q)$을 계산해야 하는 문제점이 발생한다.

<99> 따라서, $d(C_k,Q) \cong d^L(C_k,Q) \geq d^{l_k}(C_k,Q)$의 관계를 이용해 성질 1을 다음과 같이 성질1.1로 변형시킨다.

<100> 【수학식 17】

$$If\, d^{l_k}(C_k,Q)-\delta_k > d_{\min},\ \min_{X_i \in \Phi_k} d(X_i,Q) > d_{\min}$$

<101> 여기서, $l_k \leq L$

<102> 상기의 성질 1.1에 따르면, $d^{l_k}(C_k,Q)-\delta_k$이 $d_{\min}$ 보다 클 경우, 클러스터 k를 손실없이 제거할 수 있다.

<103> 반대로 $d^{l_k}(C_k,Q)-\delta_k$이 $d_{\min}$ 보다 작으면, 최적 정합자가 클러스터 k에 존재할 수 있으므로, 클러스터 k를 탐색한다. 모든 k에 대해 $d^{l_k}(C_k,Q)$과 δ_k는 이미 알려져 있기 때문에, 이러한 결정을 위한 추가적인 계산량은 없다.

<104> 이상에서와 같이, 상기 성질들을 바탕으로 출력 최적 정합자 수에 따른 두 가지 CMSA를 제시한다.

<105> 첫째는 하나의 최적 정합자를 출력하는 $CMSA_S$이며, 둘째는 다수의 상위 최적 정합자들을 출력하는 $CMSA_M$이다.

<106> 상기한 $CMSA_S$ 는 크게 세 단계로 구성된다. 먼저, MSA_S를 이용하여 $C_{k_{min}}$를 찾는다. 그리고나서, 초기 d_{min}을 $\Phi_{k_{min}}$에서 구한다. 마지막으로, 상기 성질 1.1에 의해 선택된 후보 클러스터들에 대해서만 MSA_S을 다시 적용하여 최적 정합자를 찾는다. 상기 $CMSA_S$의 탐색 과정을 요약하면 다음과 같다.

<107> 단계 1) MSA_S 를 수행하여, 최소 거리 d'_{min}를 갖는 클러스터 k_{min}을 찾는다.

<108> 단계 2) 초기 d_{min}을 d'_{min}라고 하고, MSA_S를 $\Phi_{k_{min}}$에 적용하여 d_{min}을 갱신한다.

<109> 【수학식 18】

$$d_{min} = \min_{X_i \in k_{min}} d^L(X_i, Q)$$

<110> 단계 3-1) k를 1로 한다.

<111> 단계 3-2) $k = k_{min}$이면, k를 $k+1$로 갱신한다. k>K이면, 단계 3-4으로 간다.

<112> 단계 3-3) $d'_k(C_k, Q) - \delta_k$이 d_{min} 보다 크면, 클러스터 k를 제거한다. 그렇지 않으면, MSA_S 를 Φ_k에 적용하여, d_{min}을 갱신한다. k를 $k+1$로 갱신한 후, 단계 3-2로 간다.

<113> 단계 3-4) 최종 d_{min}를 갖는 데이터를 최적 정합자로 선택한다.

<114> 상기의 $CMSA_M$의 경우도 $CMSA_S$과 동일한 방법으로 $C_{k_{min}}$을 먼저 찾는다. 그리고, 상

위 M개의 최적 정합자들의 거리 값들을 저장하기 위한 도면 4의 배열을 다음 규칙에 따라 채운다.

<115> 도면 4는 본 발명에 따라 상위 M개의 최적 정합자들의 최소 거리 배열을 나타낸 모식도이다.

<116> 먼저, $n(\Phi_{k_{min}}) \geq M$이면, $\Phi_{k_{min}}$내의 상위 M 최적 정합자들을 작은 값 순서로 배열에 채운다.

<117> 만약, $n(\Phi_{k_{min}}) < M$이면, $\Phi_{k_{min}}$내 모든 원소의 거리들을 계산하고, 값이 작은 순서로 배열에 저장한다. 남은 배열 값들에는 무한대 값을 저장한다. MSA_S를 수정하여, $\Phi_{k_{min}}$내의 상위 M 최적 정합자들을 찾을 수 있다. 이 수정된 기법을 MSA_M라고 하고, 다음과 같이 요약한다.

<118> 단계 1) 문의자 Q의 다 해상도 특징을 구한다.

<119> 단계 2) $d_{min}[\cdot]$안의 모든 원소들을 무한대 값으로 초기화한다.

<120> 단계 3) i 과 l 을 모두 1로 한다.

<121> 단계 4) $l = L$이면, 단계 6으로 간다. $i > n(\Phi_{k_{min}})$이면, 단계 7로 간다.

<122> 단계 5) $d^l(X_i, Q)$을 계산한다. $d^l(X_i, Q) > d_{min}[M-1]$이면, 현재 후보 X_i를 제거하고, i과 l를 각각 i+1과 1로 갱신한 후 단계 3으로 간다. 그렇지 않으면, l을 l+1로 갱신한 후 단계 3으로 간다.

<123> 단계 6) $d^L(X_i, Q) > d_{min}[M-1]$이면, 현재 후보 X_i를 제거한다. 그렇지 않으면, $d_{min}[M-1]$을 $d^L(X_i, Q)$로 갱신한다. 그리고, $d_{min}[\cdot]$을 작은 값 순서로 정렬한다.

i과 l을 각각 i+1과 1로 갱신한 후 단계 4로 간다.

<124> 단계 7) 최종적으로 $d_{\min}[\cdot]$에 남은 M 개의 데이터들을 최상위 M 최적 정합자들로 선택한다.

<125> 이상에서와 같이, MSA_M를 $\Phi_{k_{\min}}$에 적용하여 $d_{\min}[\cdot]$을 채운 후, 나머지 클러스터들 중 성질 1.1에 따라 선택된 각 클러스터에 MSA_M을 적용하는 방식으로 $d_{\min}[\cdot]$을 갱신한다.

<126> 최종적으로, $d_{\min}[\cdot]$에 대응하는 데이터들을 상위 M 최적 정합자들로 선택한다. 그러나, 이 탐색 기법을 통해 실제 상위 M 최적 정합자들을 정확하게 찾지 못할 경우도 있다.

<127> 도 5는 본 발명에 따라 최적 정합자들을 정확하게 못하게 되는 그릇된 판정의 예를 설명하기 위한 도면이다.

<128> 도면 5에서, X_8, X_4, X_2가 상위 3개의 최적 정합자들로 선택되었지만, 실제 3번째 최적 정합자는 X_2가 아니라 X_9 이다. 따라서, 다음과 같이 $d_{\min}$를 $d_{\min}[M-1]$로 대치함으로써 성질 1.1에서 완화된 클러스터 제거 조건인 아래의 성질 1.2를 유도할 수 있다.

<129> 【수학식 19】

$$If\ d'_k(C_k,Q)-\delta_k > d_{\min}[M-1],\ \min_{X_i\in\Phi_k} d(X_i,Q) > d_{\min}[M-1]$$

<130> 상기와 같은 성질을 본 발명이 제시하는 기법의 후처리 과정에 이용함으로써 항상 상위 M최적 정합자들을 정확하게 찾을 수 있다. 상기 성질들을 이용한 최종적인 $CMSA_M$를 요약하면 다음과 같다.

<131> 단계 1) $CMSA_S$의 단계 1과 같이, 최소 거리 $d'_{\min}$을 갖는 클러스터 $k_{\min}$을 찾는다.

<132> 단계 2) $n(\Phi_{k_{min}}) \geq M$이면, MSA_M으로 상위 M최적 정합자들을 찾아 그 거리값들을 $d_{min}[\cdot]$에 저장한다. $n(\Phi_{k_{min}}) < M$이면, $n(\Phi_{k_{min}})$개의 거리값들이 작은 값 순서로 $d_{min}[\cdot]$에 채워지며, 나머지는 무한대값으로 채운다.

<133> 단계 3-1) k를 1로 한다.

<134> 단계 3-2) $k = k_{min}$이면, k 를 k + 1로 갱신한다. k>K이면, 단계 3-5으로 간다.

<135> 단계 3-3) $d^{l_k}(C_k, Q) - \delta_k > d_{min}[0]$ 이면, 클러스터 k를 제거하고, k를 k+1로 갱신한 후 단계 3-2로 간다.

<136> 단계 3-4) MSA_M을 Φ_k에 적용함으로써 $d_{min}[\cdot]$을 갱신한다. k를 k+1로 갱신한 후 단계 3-2로 간다.

<137> 단계 3-5) k를 1로 한다.

<138> 단계 3-6) 클러스터 k가 단계 3-4에서 이미 조사되었으면, k 를 k + 1로 갱신한다. $k > K$이면, 단계 3-9로 간다.

<139> 단계 3-7) $d^{l_k}(C_k, Q) - \delta_k > d_{min}[M-1]$이면, 클러스터 k를 제거하고 k를 k+1로 갱신한 후 단계 3-6으로 간다.

<140> 단계 3-8) MSA_M를 Φ_k에 적용함으로써 $d_{min}[\cdot]$를 갱신한다. k를 k+1로 갱신한 후 단계 3-6으로 간다.

<141> 단계 3-9) 최종 $d_{min}[\cdot]$에 대응하는 M개의 데이터들을 최적 정합자들로 선택한다.

【발명의 효과】

<142> 이상에서와 같이 본 발명에 의한 대용량 데이터베이스에서의 고속에 의한 다해상도

의 최적 탐색방법에 따르면 다음과 같은 이점이 있다.

<143> 첫째, 대용량 데이터베이스에서의 고속 최적 탐색을 위한 모든 시스템 즉, 영상, 동영상 데이터베이스에 대한 탐색 엔진의 핵심 모듈로 사용할 수 있다.

<144> 둘째, 영상이나 음성 등의 다 해상도 구조가 가능한 모든 멀티미디어 데이터베이스에 적용시킴으로써 매우 빠른 속도로 원하는 정보를 데이터베이스에서 정확하고 빠르게 찾을 수 있다.

【특허청구범위】

【청구항 1】

다 해상도 데이터베이스 기반에서의 고속 탐색방법에 있어서,

문의자 특징 Q의 다 해상도 구조를 구하는 단계와;

초기 d_{min}을 무한대로 설정하는 단계와;

i 와 l를 모두 1로 설정하는 단계와;

$d^l(X_i,Q)$을 구하는 단계와;

$d^L(X_i,Q)$을 구하는 단계와;

최종 d_{min}을 갖는 데이터를 최적 정합자로 선택하는 단계로 이루어지는 대용량 데이터베이스에서의 고속에 의한 다해상도의 최적 탐색방법.

【청구항 2】

청구항 1에 있어서, 상기 $d^l(X_i,Q)$을 구하는 단계에서 상기 $d^l(X_i,Q)$값이 d_{min} 보다 크면 현재 후보 X_i를 제거함과 더불어 i와 l 을 각각 i+1과 1로 갱신하고, 그렇지 않으면 l을 l+1로 갱신하는 과정을 거치는 것을 특징으로 하는 대용량 데이터베이스에서의 고속에 의한 다해상도의 최적 탐색방법.

【청구항 3】

청구항 1에 있어서, 상기 $d^L(X_i,Q)$을 구하는 단계에서 상기 $d^L(X_i,Q)$값이 d_{min} 보다 크면 현재 후보 X_i를 제거하고, 그렇지 않으면 d_{min}을 $d^L(X_i,Q)$로 갱신함과 더

불어 i와 l을 각각 i+1과 1로 갱신하는 과정을 거치는 것을 특징으로 하는 대용량 데이터베이스에서의 고속에 의한 다해상도의 최적 탐색방법.

【청구항 4】

청구항 1에 있어서, 상기 다 해상도 데이터 베이스에서의 고속 탐색은 다음과 같은 성질의 부등식에 의해 유도되는 것을 특징으로 하는 대용량 데이터베이스에서의 고속에 의한 다해상도의 최적 탐색방법.

【수학식 20】

$$d(X,Y) \cong d^{L}(X,Y) \geq d^{L-1}(X,Y) \geq \cdots \geq d^{l}(X,Y) \geq \cdots \geq d^{l}(X,Y) \geq d^{0}(X,Y)$$

$d^{l}(X,Y)$: 계층 l에서의 두 히스토그램 X와 Y의 L_1-놈 거리

【청구항 5】

클러스터 기반에서 하나의 최적 정합자를 출력하는 $CMSA_S$방식을 이용한 다 해상도 탐색 방법에 있어서,

고속 다 해상도 탐색방법(MSA_S)을 수행하여 최소 거리 d'_{min}를 갖는 클러스터 k_{min}을 찾는 단계와;

초기 d_{min}을 d'_{min}로 하고, 상기 MSA_S를 $\Phi_{k_{min}}$에 적용하여 d_{min}을 갱신하는 단계와;

$d^{l_k}(C_k,Q)-\delta_k$값을 구하는 단계와;

최종 d_{min}를 갖는 데이터를 최적 정합자로 선택하는 단계로 이루어지는 대용량 데이터베이스에서의 고속에 의한 다해상도의 최적 탐색방법.

【청구항 6】

청구항 5에 있어서, 상기 $CMSA_S$방식을 이용한 다 해상도 탐색은 다음과 같은 성질의 부등식에 의해 유도되는 것을 특징으로 하는 대용량 데이터베이스에서의 고속에 의한 다해상도의 최적 탐색방법.

【수학식 21】

$$If\, d^{l_k}(C_k,Q)-\delta_k > d_{\min},\ \min_{X_i \in \Phi_k} d(X_i,Q) > d_{\min}$$

단, $l_k \le L$

【청구항 7】

청구항 5에 있어서, 상기 $d_{\min}$은

【수학식 22】

$$d_{\min} = \min_{X_i \in k_{\min}} d^{L}(X_i,Q)$$

k는 1로 하며, $k = k_{\min}$이면 k를 $k+1$로 갱신하는 것을 특징으로 하는 대용량 데이터베이스에서 고속에 의한 다해상도의 최적 탐색방법.

【청구항 8】

청구항 5 또는 청구항 6에 있어서, 상기 $d^{l_k}(C_k,Q)-\delta_k$값을 구하는 단계에서 $d^{l_k}(C_k,Q)-\delta_k$값이 $d_{\min}$ 보다 크면 클러스터 k를 제거하고, 그렇지 않으면 MSA_S 를 Φ_k에 적용하여, $d_{\min}$을 갱신함과 더불어 k를 $k+1$로 갱신하는 것을 특징으로 하는 대용량 데이터베이스에서의 고속에 의한 다해상도의 최적 탐색방법.

【청구항 9】

클러스터 기반에서 다수의 상위 최적 정합자를 출력하는 $CMSA_M$방식을 이용한 다 해상도 탐색 방법에 있어서,

고속 다 해상도 탐색방법(MSA_S)을 수행하여 최소 거리 d'_{min}를 갖는 클러스터 k_{min}을 찾는 단계와;

$n(\Phi_{k_{min}}) \geq M$이면, MSA_M(MSA_S를 수정하여 $\Phi_{k_{min}}$내의 상위 M 최적 정합자들을 찾을 수 있도록 수정된 기법)으로 상위 M최적 정합자들을 찾아 그 거리값들을 $d_{min}[\cdot]$에 저장하는 단계와;

k 를 1로 설정 하여 $k = k_{min}$이면, k 를 $k + 1$로 갱신하는 단계와;

$d'_k(C_k, Q) - \delta_k > d_{min}[0]$ 이면, 클러스터 k를 제거하고 k를 k+1로 갱신하는 단계와;

MSA_M을 Φ_k에 적용하여 $d_{min}[\cdot]$을 갱신하고, k를 k+1로 갱신하는 단계와;

k를 1로 설정하여 클러스터 k가 이미 조사되었으면, k 를 $k + 1$로 갱신하는 단계와;

$d'_k(C_k, Q) - \delta_k > d_{min}[M-1]$이면, 클러스터 k를 제거하고, k를 k+1로 갱신하는 단계와;

MSA_M를 Φ_k에 적용하여 $d_{min}[\cdot]$를 갱신하고, k를 k+1로 갱신하는 단계와;

최종 $d_{min}[\cdot]$에 대응하는 M개의 데이터들을 최적 정합자들로 선택하는 단계로 이루어지는 대용량 데이터베이스에서의 고속에 의한 다해상도의 최적 탐색방법.

【청구항 10】

㈜ 청구항 9에 있어서, 상기 $CMSA_M$방식을 이용한 다 해상도 탐색은 다음과 같은 성질의 부등식에 의해 유도되는 것을 특징으로 하는 대용량 데이터베이스에서의 고속에 의한 다해상도의 최적 탐색방법.

【수학식 23】

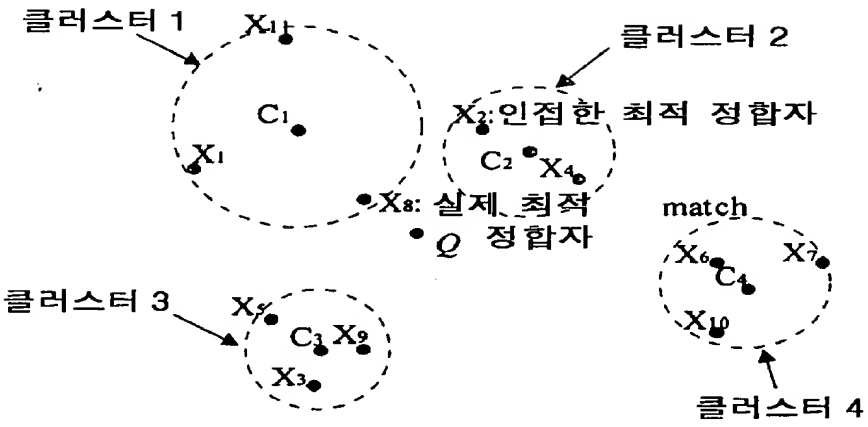
$$If\, d^{l_k}(C_k, Q) - \delta_k > d_{\min}[M-1],\ \min_{X_i \in \Phi_k} d(X_i, Q) > d_{\min}[M-1]$$

【청구항 11】

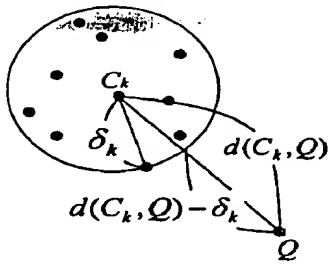
청구항 9에 있어서, 상기 $n(\Phi_{k_{\min}}) < M$이면, $n(\Phi_{k_{\min}})$개의 거리값들이 작은 값 순서로 $d_{\min}[\cdot]$에 채워지며, 나머지는 무한대값으로 채워지는 것을 특징으로 하는 대용량 데이터베이스에서의 고속에 의한 다해상도의 최적 탐색방법.

【도면】

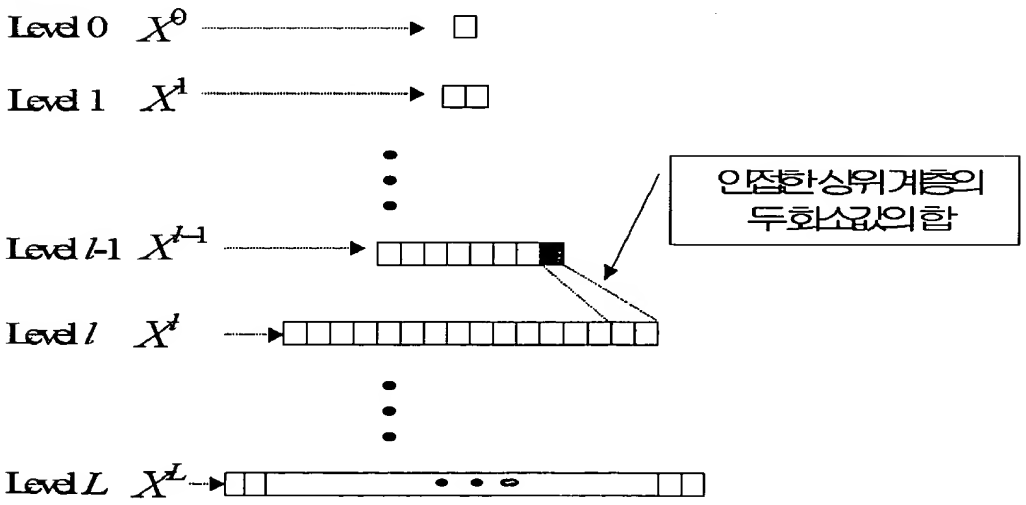
【도 1】


클러스터 1
클러스터 2
C1
X1
X2:인접한 최적 정합자
C2
X4
X8: 실제 최적
Q 정합자
match
X6
X7
C4
X10
클러스터 3
C3
X2
X3
클러스터 4

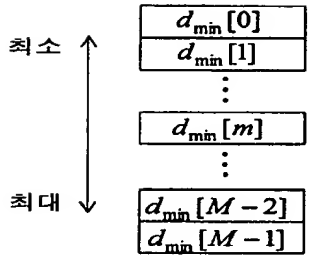
【도 2】


C_k
δ_k
$d(C_k,Q)$
$d(C_k,Q)-\delta_k$
Q

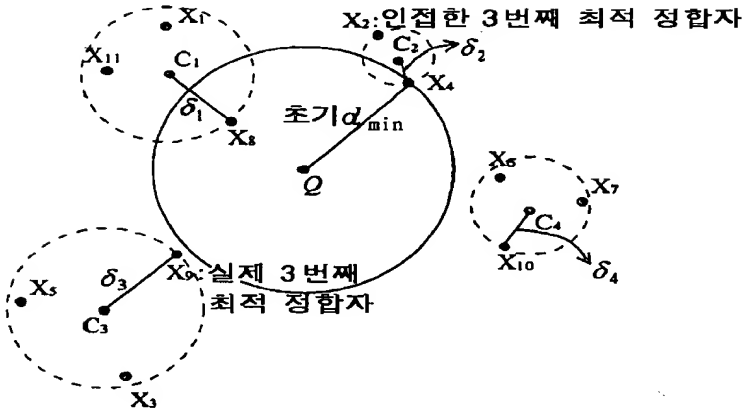
【도 3】


Level 0 X^0
Level 1 X^1
Level l-1 X^{l-1}
Level l X^l
Level L X^L
인접한상위계층의
두화소값의합

【도 4】


최소
최대
$d_{min}[0]$
$d_{min}[1]$
$d_{min}[m]$
$d_{min}[M-2]$
$d_{min}[M-1]$

【도 5】


X₂:인접한 3번째 최적 정합자
초기 d_{min}
Q
X₂:실제 3번째
최적 정합자